BAREKENG: Journal of Mathematics and Its Applications
March 2026 Volume 20 Issue 1 Page 0179–0196
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol20iss1pp0179-0196

MAPPING DISASTER-PRONE AREAS ON JAVA ISLAND USING THE K-PROTOTYPES ALGORITHM

Mey Lista Tauryawati [1*], **Ahmad Fuad Zainuddin** [2]

[1,2] *Business Mathematics Department, Universitas Prasetiya Mulya*
Jln. BSD Raya Utama No. 1, Kavling Edutown, Tangerang, 15339, Indonesia

*Corresponding author's e-mail: * **mey.lista@prasetiyamulya.ac.id***

Article Info

Article History:

Received: 23rd February 2025
Revised: 2nd June 2025
Accepted: 17th July 2025
Available online: 24th November 2025

Keywords:

Clustering;
Disaster;
K-Prototypes.

ABSTRACT

Clustering in disaster areas is often implemented as a disaster mitigation effort with the aim of minimizing risk. Determining the appropriate clustering method based on the data set will influence the clustering results. K-Prototypes is a clustering method that is capable of handling mixed data, numerical and categorical data, so this method is suitable to clustering disaster prone area with mixed data of disaster factors such as incident intensity, type of disaster, population density, and level of infrastructure vulnerability. This research focuses on disaster prone areas on Java Island and clustering using K-Prototypes to group and map areas that have the highest to lowest levels of disaster vulnerability based on the number of incidents, number of victims, and the amount of damage to facilities and the type of disaster. The clustering results obtained mapping of cities in the province into cluster groups based on the level of vulnerability and calculated potential losses based on disasters in each province. Afterward, the clustering results are used to determine priority areas for disaster mitigation to minimize losses.

How to cite this article:

M. L. Tauryawati and A. F. Zainuddin., "MAPPING DISASTER-PRONE AREAS ON JAVA ISLAND USING THE K-PROTOTYPES ALGORITHM," *BAREKENG: J. Math. & App.*, vol. 20, iss. 1, pp. 0179-0196, Mar, 2026.

Journal homepage: *https://ojs3.unpatti.ac.id/index.php/barekeng/*
Journal e-mail: *barekeng.math@yahoo.com; barekeng.journal@mail.unpatti.ac.id*

Research Article · **Open Access**

1. INTRODUCTION

In the last 10 years from 2021 to 2022 the total number of natural disasters that occurred in Indonesia reached 33,080 natural disasters recorded in the data of the Indonesia's National Disaster Management Agency or BNPB. From 2015 to 2021 the number of natural disasters consistently showed a significant increase with an average increase rate of 22% or an increase in the number of disaster events reaching approximately 600 disaster events annually [1]. The losses caused by natural disasters are indeed diverse, including numerous fatalities, injuries, and cases of missing persons. In addition to casualties, natural disasters also cause material losses such as damage to facilities and infrastructure ranging from residents' homes, educational facilities, health facilities, and worship facilities. Throughout 2022, there were 3,544 natural disasters, resulting in 861 deaths, 46 missing persons, 8,272 injured, and 8 million people affected. Regarding facility damage, 95,000 houses were damaged to varying degrees, along with 1,983 public facilities. Meanwhile, in 2021, 5,402 natural disasters occurred, causing 728 deaths, 87 missing persons, 14,000 injuries, and 7 million people affected. Facility damage included 150,000 houses with varying degrees of damage, including severe destruction, and 1,400 public facilities were also damaged.

Among the many natural disasters that occurred at the same time based on data from Indonesia's National Disaster Management Agency, Java Island became one of the islands with the highest number of natural disasters in Indonesia, in 2022, as many as 1,838 natural disasters occurred on Java Island from a total of 3,544 disasters, this indicates that throughout 2022, 50% or half of the total number of natural disasters in Indonesia occurred on Java Island [1]. There are many factors that cause Java Island to be one of the areas with a high number of natural disasters in Indonesia, one of which is due to the geographical location of Java Island through which tectonic plates and volcanic rings are so active in parts of West Java, Central Java, and East Java [2].

In addition to its geographical location, the population density in Java Island is also one of the causes why Java Island is prone to disasters and environmental damage around it and affects the disaster index and high risk on the island of Java [3]. Based on this, the government should pay more attention to mitigation or disaster management in Java. Society also needs to know how the impact of natural disasters that occur on the island of Java and know which areas are prone to disasters when viewed from how often a disaster occurs in the area, the types of disasters that often occur, and how the impact of losses incurred for future anticipatory steps in minimizing the risk of casualties and property losses. Preparation and planning for disaster management or mitigation are crucial, especially on data and information that can be used as a basis for making decisions on mitigation efforts, with the aim of reducing losses caused by disasters. The information and data can be obtained through the process of analyzing and processing historical data of natural disasters that have existed before. Many methods can be used to help process and analyze data, the most popular example today is data mining, a method that has been around for a long time. Among the many data mining techniques available, clustering is one of the simplest and most frequently used by researchers. Clustering is included in the unsupervised learning, which is a machine learning that does not require an equation model but determines the pattern of data [4]. The foundation of machine learning lies in using statistical inference and mathematical modeling to analyze data, detect patterns, and perform tasks such as classification or prediction [5].

Research on clustering with the K-Prototypes method has been conducted by Annas et al for clustering tectonic earthquakes on Sulawesi Island. The use of K-Prototypes method as a clustering method is because the variables used are mixed (numerical and categorical) [6]. Another paper that discusses the use of the K-Prototype algorithm for clustering on mixed data was conducted by Refaldy et al for-clustering cities in South Sulawesi Province based on community welfare indicators. The K-Prototype algorithm was chosen because of its ability to process data with mixed types, numerical and categorical. One of the challenges in applying this algorithm is determining the optimal number of clusters. To overcome this, researchers can use the Elbow method which analyses the Sum of Squared Errors (SSE) graph against various numbers of clusters [7].

This research will be conducted with the aim of clustering the level of disaster vulnerability of regions on the Java Island with the K-Prototypes method to observe which areas are included in the vulnerable category or not based on the type of natural disaster, the number of events, the number of casualties, and the amount of damage to facilities caused by the disaster that occurred. The use of the K-Prototypes algorithm in this study presents a notable methodological contribution, as it effectively handles mixed-type data, comprising both numerical variables and categorical variables. Traditional clustering methods such as K-Means are limited to numerical data, while K-Modes is suitable only for categorical data.

In contrast, K-Prototypes integrates the strengths of both, making it particularly suitable for real-world disaster datasets that typically contain a combination of variable types. The results of clustering areas based on natural disaster data will serve as a reference or guideline to identify disaster-prone areas by considering the intensity of the incident, the amount of loss caused, and the types of disasters occurring in the area. Additionally, the analysis of casualties and material losses can help assess the risks associated with the region's level of natural disaster vulnerability.

2. RESEARCH METHODS

This research uses a disaster dataset taken from BNPB which occurred in Java Island in a period of 10 years, starting from 2012 to 2022 with data attributes that will be used, namely the Name of Provinces in Java Island, Number of Events Per Province, Number of Victims, and Number of Damaged Facilities. The disaster data includes natural disaster data categories, namely landslides, earthquakes, floods, Tornadoes, and forest and land fires. In this research, the mapping of disaster-prone areas per regency in the province of Java Island was carried out. Furthermore, the clustering results serve as a basis for estimating future potential economic losses, as each cluster represents regions with similar historical disaster profiles. Potential loss refers to the estimated economic damage reported for each disaster event, measured in Indonesian Rupiah (IDR). The data was obtained from official records published by BNPB between 2012 and 2022. These values include infrastructure damage, property loss and other direct economic impacts.

This research aims to apply K-Prototypes in mapping disaster-prone areas in Java Island based on the intensity of events and types of disasters. The results of this mapping will be analyzed to determine the relationship between the impact of losses incurred with the level of disaster vulnerability in each group of areas. The impact of loss analysis can be seen from the calculation of potential loss analysis per cluster for each type of disaster. The results of this study can be used as reference material for the government or related institutions in determining preventive efforts to minimize the impact of natural disaster losses, especially in Java.

Basically, natural disasters can be grouped from the causes of occurrence as follows: first, geological natural disasters, which are disasters that occur due to the movements of elements in the bowels of the earth such as the movement of tectonic plates that can result in earthquakes and tsunamis. Second, climatological natural disasters, which are disasters that can occur due to changes or climatic conditions such as floods, strong winds, and drying springs. And third, extra-terrestrial natural disasters, which are disasters caused by activities in the solar system such as the movement of asteroids and meteors colliding in the sky that can have an impact on the earth's surface.

2.1. Clustering Method

Data mining is a process of discovering new relationships, patterns and habits in data or objects. This process is done by extracting big data and processing or analyzing it using mathematical processes. Data mining can be a combination of machine learning techniques with the application of mathematical or statistical sciences such as pattern recognition, statistical processing, and visualization to handle information retrieval problems from large databases [8], [9]. Data clustering is a process to group datasets whose class attributes have not been described. A simple explanation of clustering is that if we have a large data set, the first process we do is to group the large data into several clusters and then transform it into an ordered set so that it can be simplified into certain groups. Cluster can also be interpreted as a collection of things. Before doing the analysis, it is necessary to understand that a given dataset already has similarities between its members. Therefore, each member with the same characteristics is divided into several groups [10]. The purpose of clustering this data is to minimize diversity within groups and maximize the type of diversity between groups [11].

The clustering method is part of unsupervised data mining because it does not analyze the relationship between variables. This technique is used to group data that has similarities in a large database. This clustering also makes it easier for us to analyze large data into small groups so that it is more effective and efficient. Clustering can be divided into two, the first is hierarchical clustering, which classifies data into clusters with a clear hierarchy (level) between data. The second is the non-hierarchical clustering method, which is clustering without a hierarchy by randomly determining the number of clusters at the beginning. One example of non-hierarchical clustering is K-Means Clustering and K-Prototypes [4], [11].

K-Means Clustering is one of the non-hierarchical clustering methods where the clustering is done by determining the initial number of clusters by randomly determining the initial centroid value. In K-Means, data is partitioned into separate sets $\boldsymbol{C_1, C_2, \dots, C_k}$ with $\boldsymbol{C_i}$ each represented by a centroid or center point. K-Means Clustering will work by measuring the squared distance between each point $\boldsymbol{X}$ to the center of mass of its cluster or called the centroid, which will be grouped based on the results of measuring the smallest distance to the centroid. The smaller the distance of the point to the centroid means that the data collected is closer together. In other words, minimizing the Within-Cluster Sum of Squares (WCSS) value which is defined in Eq. (1) as follows:

$$\boldsymbol{\arg\min \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|^2 ,} \tag{1}$$

$\boldsymbol{\mu_i}$ is the means or average value of the points in cluster $\boldsymbol{C_i}$ which is also known as the centroid. The centroid value in K-Means can be defined in Eq. (2) as follows:

$$\boldsymbol{\mu_i(C_i) = \arg\min \sum_{x \in C_i} d(x, \mu)^2} \tag{2}$$

In K-Means Clustering the centroid starting point will be determined randomly. Data that is close to the center or centroid will form a new data group or group. The process continues to repeat or is called iterative until there is no change in each group or it has converged [12] . Research on clustering disaster areas has also been carried out by Dewi et al using K-means clustering [13], but the limitations of using the K-means clustering method can only be used for numerical data.

K-Prototypes is one of the partition-based clustering methods. This method is the result of the development of the K-Means method to handle clustering on numerical data and K-Modes for categorical data. The smaller the distance of a point to the centroid means that the clustered data is closer together. The general steps to use the K-Prototypes algorithm in clustering disaster-prone areas are:

1. Data preparation
   Collect relevant data on disaster-prone areas, including numerical and categorical data. In addition, before performing data processing, it is necessary to ensure that the data is ready to be processed and ready to be used.
2. Normalization of numerical data
   Perform normalization on numerical data to avoid bias caused by different scales. Commonly used normalization methods are min-max scaling or $\boldsymbol{z}$-score normalization.
3. Encoding categorical data
   Convert categorical data into a numerical representation that can be used in clustering algorithms. You can use methods such as one-hot encoding or label encoding, depending on the data characteristics and analysis needs.
4. Determine the number of clusters (*k*)
   Determine the desired number of clusters to group disaster-prone areas. This can be done based on domain knowledge, visual analysis, or using appropriate cluster number selection methods such as elbow method or silhouette analysis.
5. Centroid initialization
   Perform initial initialization for the centroid or cluster center using a suitable method. A common initializations option is to randomly select some data points as the initial centroid.
6. Clustering iteration
   Running the k-prototypes algorithm by performing iterations to optimize the centroid position and group the data into appropriate clusters. This process involves calculating the distance between the data and the centroid and updating the centroid position based on the data.
7. Evaluation of clustering results
   Evaluate the clustering results using evaluation metrics. Visual analysis can also help in understanding and interpreting the clustering results.
8. Interpretation and follow-up
   Interpret and analyze the clustering results to gain insight into the patterns and characteristics of the disaster-prone areas formed in each cluster.

2.2. Evaluation Method

Silhouette coefficient is an evaluation metric used to measure the quality of clustering or separation of data into groups (clusters). The silhouette coefficient provides information on how well each data object fits in the cluster it is placed in. The silhouette coefficient combines two concepts, namely cohesion and separation. Cohesion describes how close the data objects in a cluster are to other data objects in the same cluster. Separation describes how far data objects in a cluster are from data objects in other clusters [14]. It is necessary to calculate the silhouette index value of the $\boldsymbol{i}$-th data to calculate the silhouette coefficient value. The silhouette coefficient value is obtained by finding the maximum value of the Global Silhouette Index value from cluster number 2 to cluster number $\boldsymbol{n-1}$ with the following Eq. (3):

$$\boldsymbol{SC = (\max)_{k}\, SI(k)} \tag{3}$$

where $\boldsymbol{SC}$ is the Silhouette Coefficient, SI is the Global Silhouette Index, and $\boldsymbol{k}$ is the number of clusters. To calculate the $\boldsymbol{SI}$ value of an $\boldsymbol{i}$-th data, there are 2 components, namely $\boldsymbol{a_i}$ and $\boldsymbol{b_i}$. The value of $\boldsymbol{a_i}$ is the average of the $\boldsymbol{i}$-th distance to all other data in one cluster, while $\boldsymbol{b_i}$ is obtained by calculating the average distance of the $\boldsymbol{i}$-th data to all data from other clusters that are not in the same cluster as the $\boldsymbol{i}$-th data, then the smallest is taken [11]. The equations to calculate the values of $\boldsymbol{a_i^j}$ and $\boldsymbol{b_i^j}$ are defined in Eqs. (4) and (5) as follows:

$$\boldsymbol{a_i^j = \frac{1}{m_j}\sum_{\substack{r=1\\ r\neq 1}}^{m_j} d\left(x_i^j, x_r^j\right),} \tag{4}$$

$$\boldsymbol{b_i^j = (\min)_{\substack{n=1,2,\dots,k\\ n\neq j}}\left\{\frac{1}{m_n}\sum_{\substack{r=1\\ r\neq 1}}^{m_n} d\left(x_i^j, x_r^n\right)\right\}.} \tag{5}$$

From Eqs. (4) and (5), the equations to obtain the values of SI_i^j, SI_j, and SI are shown in equations Eqs. (6) - (8) as follows:

$$\boldsymbol{SI_i^j = \frac{b_i^j - a_i^j}{\max\left\{a_i^j, b_i^j\right\}},} \tag{6}$$

$$\boldsymbol{SI_j = \frac{1}{m_j}\sum_{i=1}^{m_j} SI_i^j,} \tag{7}$$

$$\boldsymbol{SI = \frac{1}{k}\sum_{j=1}^{k} SI_j,} \tag{8}$$

with the description:

$\boldsymbol{j}$	: $\boldsymbol{j}$-th cluster;
$\boldsymbol{i}$	: $\boldsymbol{i}$-th data ($\boldsymbol{i = 1, 2, \cdots, m_j}$);
$\boldsymbol{a_i^j}$	: average distance of i-th data to all data in one;
cluster $\boldsymbol{m_j}$	: number of data in $\boldsymbol{j}$-th;
cluster $\boldsymbol{d\left(x_i^j, x_r^j\right)}$	: distance of $\boldsymbol{i}$-th data to $\boldsymbol{r}$-th data in one cluster $\boldsymbol{j}$;
b_i^j	: the average distance of the i-th data to all data that is not in a cluster;
m_n	: number of data in the n-th cluster;
$d\left(x_i^j, x_r^n\right)$	: distance of i-th data to j-th data in one cluster-n;
SI_i^j	: silhouette index of i-th data in a cluster;
SI_j	: silhouette index of-i ($i = 1, 2, \dots, m_j$) in a cluster;
k	: number of cluster data.

The subjective criteria for clustering measurement based on the Silhouette Coefficient according to Kauffman and Roesseeuw [14], [15] can be seen in Table 1 as follows:

Table 1. Silhouette Measurement Criteria

Silhouette Coefficient Value	Criteria
0.71 – 1.00	Extremely Robust Structure
0.51 – 0.70	Robust Structure
0.26 – 0.50	Weak Structure
≤ 0.25	Extremely Weak Structure

Table 1 illustrates the classification of silhouette measurement criteria into several categories based on the silhouette score values. The silhouette score ranges from -1 to 1, where a positive value indicates that an object is appropriately assigned to its cluster, whereas a negative value suggests that the object may have been misclassified. A silhouette score approaching 1 signifies that the object is well matched to its own cluster and poorly matched to neighbouring clusters. A score near 0 indicates that the object is located on or very close to the decision boundary between two adjacent clusters, potentially implying cluster overlap. A negative score indicates that the object may have been assigned to an incorrect cluster. Specifically, a silhouette score in the range of 0.71 to 1.00 denotes an excellent fit within the cluster. Scores between 0.51 and 0.70 represent a good fit, while scores from 0.26 to 0.50 reflect suboptimal clustering. Scores below 0.25 suggest poorly defined cluster boundaries, with many objects situated near or between neighbouring clusters [16].

The Davis-Bouldin Index is a method used to measure the extent to which a clustering algorithm is successful in separating distinct groups (clusters). This index combines measures of density and distance between the clusters formed. The Davies Bouldin (DB) Index checks the sum of the similarity of the data to the cluster centers of the cluster and the distance between the cluster centers of the cluster [17]. The DB index produces a numerical value that represents the quality of the clustering. The lower the DB index value, the better the clustering algorithm is at separating different groups. The minimum DB index value is zero, which indicates perfect clustering. The DB index is calculated by finding the Sum of Square Within (SSW) value of the cluster to determine the distance of points in an i-th cluster which can be seen in Eq. (9) as follows [18]:

$$SSW_i = \frac{1}{m_i}\sum_{j=i}^{m_i} d\left(x_j, c_i\right), \tag{9}$$

with m_i as the number of data in i-th cluster; c_i as the i-th centroid cluster; and $d(x_j, c_i)$ as the Euclidian distance in each data to centroid.

2.3. Data

The data used in this research is secondary data consisting of a dataset of natural disasters that occurred in Java Island within a period of 10 years, namely 2012 - 2022. The natural disaster data used consists of 5 (five) natural disaster variables, namely Landslides, Earthquakes, Floods, Tornadoes, and Forest and Land Fires. All data used were taken from Information Data of Indonesia's National Disaster Management Agency (DIBI BNPB). The initial stages of the work began with data collection taken from the Indonesian Disaster Information Data website of BNPB. The data taken are data on the types of disasters Floods, Landslides, Tornadoes and Earthquakes, data on the names of provinces in Java Island there are 6 (six) namely Banten, DKI Jakarta, West Java, Central Java, East Java, and D.I Yogyakarta, data on Cities in Java Island there are 119 Cities, Number of Disaster Events, Number of Casualties, and Number of Damage to Facilities. The Natural Disaster dataset can be seen in Table 2 below:

Table 2. Natural Disasters Dataset

Disaster Types	Province	Regency	Disaster Count (Events)	Number of Casualties (People)	Number of Facility Damages (Unit)
Flood	Banten	Lebak	37	68,967	382
Flood	Banten	Pandeglang	25	260,132	270
Flood	Banten	Serang	81	131,013	433
Flood	Banten	Tangerang	17	255,556	20
Flood	Banten	Cilegon	15	92,096	10
⋮	⋮	⋮	⋮	⋮	⋮
Earthquake	East Java	Surabaya	0	0	0
Earthquake	East Java	Batu	1	0	8

The next stage is data pre-processing; this stage is carried out to prepare and process raw data into data that is ready for analysis or mining process. At this stage, checking for missing values or empty values in the data and checking outlier data. In the data on the number of events, number of victims, and number of damaged facilities in all categories of natural disasters have several outlier values and some of the values are quite extreme. To minimise the extremity of these outliers, outliers will be handled using winsorization technique by replacing the outlier values with the nearest less extreme values [19]. Winsorization can be done by selecting a certain percentile boundary, such as the 1st percentile or 99th percentile, as the cutting or replacement point. Values below or above the percentile limit will be replaced with the percentile value. In the data used in this study, extreme outlier values will be replaced with the lower limit value of the 5% percentile and the upper limit of the 95% percentile. After winsorization, the outlier data that was so extreme before has reduced the level of extremity [19], [20].

The next stage is data standardisation. Data standardisation is the process of converting data into a uniform scale or having consistent characteristics. The aim is to facilitate comparison and analysis of data that have different scales or have large variations. With standardisation, variables with different scales or large variability can be aligned, allowing for more accurate decision-making and the discovery of clearer patterns in the data. At this stage, the numerical data that has gone through the outlier handling process will be scaled. The scale function will standardise the data using the Z-score method. The Z-score method, also known as standard score, is used to transform data by turning it into a normal distribution with a mean of 0 and a standard deviation of 1. In the context of the scale function, each value in the dataset will be reduced by the mean of the dataset, then the result will be divided by the standard deviation of the dataset. Thus, the data will be on the same scale and can be compared relatively.

3. RESULTS AND DISCUSSION

This research presents the results of clustering of disaster-prone areas on the island of Java based on the number of incidents, number of victims, and amount of damage to facilities using the K-Prototypes Method. The data was taken from information data of BNPB with the types of disasters observed are: Floods, Landslides, Tornadoes and Earthquakes. The province observed in this research include Banten, DKI Jakarta, West Java, Central Java, Java East, and D.I Yogyakarta and the data consist of 119 cities across Java Island, including information on the number of disaster events, casualties, and damages facilities.

3.1 Clustering of Earthquake-Prone Areas

The clustering of earthquake-prone areas aims to categorize regions based on their vulnerability to earthquake occurrences. Table 4 Clustering of earthquake-prone areas is determined from cluster visualization by determining a random value k which states the number of clusters. The value of k can be set to $k = 2, 3, 4, 5, \dots, n$. Afterwards, cluster evaluation will be carried out with Davis Bouldin Index (DBI) and a lower DBI indicates better separation between clusters and higher clustering quality. From the calculation of the Davies Bouldin Index, it can be concluded that the number of clusters of earthquake-prone areas is 2 clusters.

Table 3. Davis Bouldin Index of Earthquake

Number of Cluster	Davis Bouldin Index
2	1.218
3	2.153
4	2.074
5	1.904

Table 4. Clustering Earthquake Prone Areas

Cluster	Number of Incidents	Number of Casualties	Number of Damage to Facilities
0	2	1	37
1	33	2104	2418

The frequency of earthquake disasters, casualties, and damage caused in cluster 0 is very small compared to cluster 1. This indicates that cluster 1 has a greater risk and impact from earthquake disasters compared to cluster 0. Cluster 0 describes earthquake disasters that tend to have lower values in terms of the number of occurrences, the number of fatalities, and the level of damage to facilities. Most regencies in the sample fall into this cluster, with 102 out of a total of 118 regencies (approximately 86.4%). The most common regency is East Java province, with 35 out of 38 regencies in the sample. However, it should be noted that the number of damaged facilities in this province is very high, reaching 6277 units. Cluster 1 describes earthquake disasters that tend to have higher values in terms of the number of events, the number of casualties, and the level of damage to facilities. The number of samples in this cluster is 16 regencies (approximately 13.6%). West Java is the most province affected by earthquakes, with 9 regencies in this cluster. However, the worst damage occurred in East Java Province, where the number of damaged facilities reached 19,192 units. The frequency of earthquake disasters, casualties and damage occurring in cluster 0 is very small compared to cluster 1. This indicates that cluster 1 has a greater risk and impact of earthquake disasters compared to cluster 0.

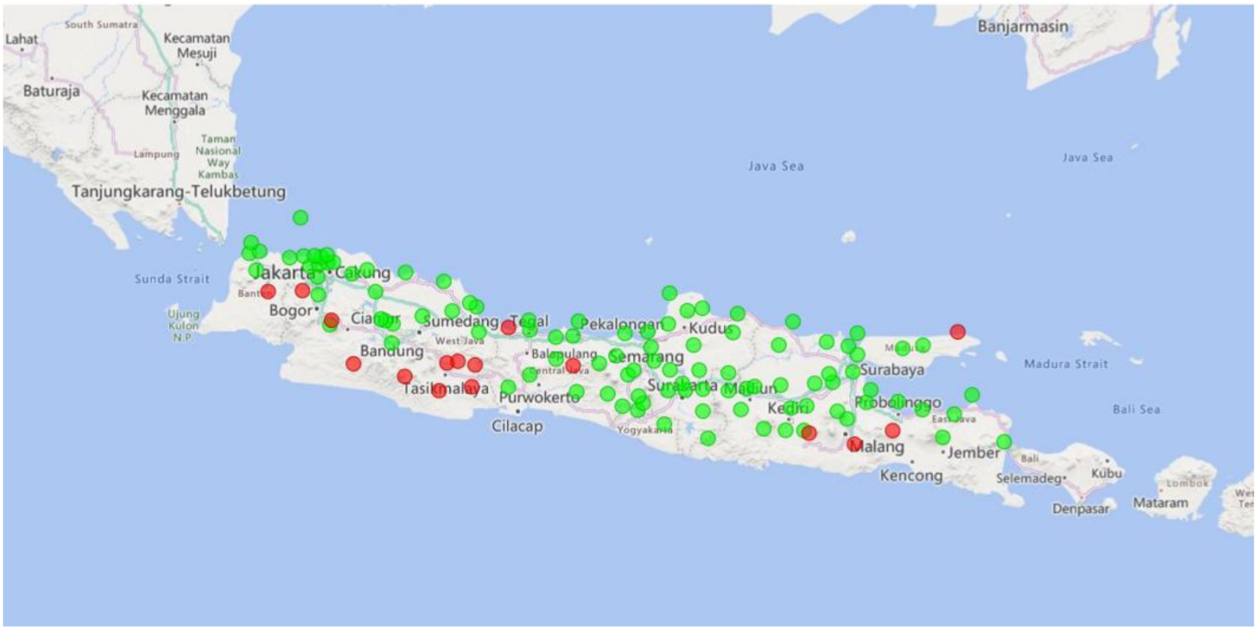

Figure 1. Figure Clustering of Earthquake Prone Areas

In Fig. 1, red dots represent areas with high risk (cluster 1), while green dots indicate low risk (cluster 0). The distribution of earthquake-prone areas is dominated by West Java and some areas of Banten, as well as East Java. The affected areas are in coastal areas facing the Indian Ocean. It is likely that the earthquake occurred because the area coincides with the location where the Indo-Australian and Eurasian Plates meet, making it vulnerable to earthquakes. These areas include the southern region of West Java, then Banten, especially areas along the South Coast, several areas in Central Java such as Yogyakarta, Semarang, and Solo are also within the earthquake-prone zone, and several areas known to have a high earthquake risk in East Java include Surabaya, Malang, and surrounding areas.

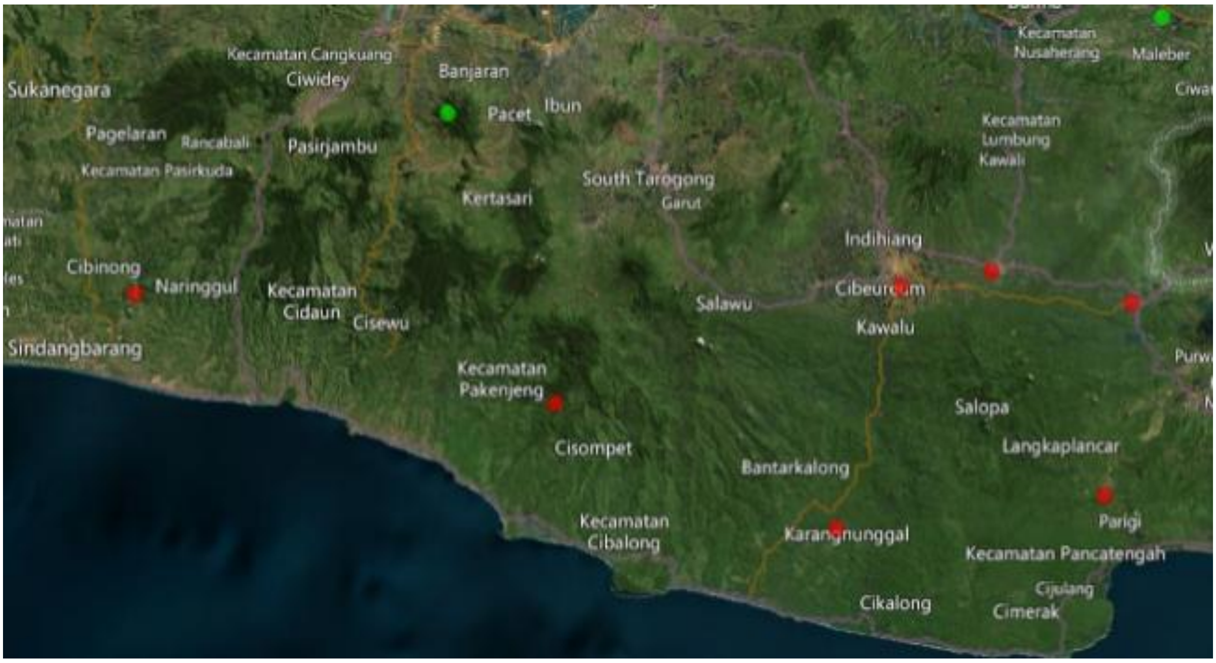

Figure 2. The Zoomed in Clustering of Earthquake Prone Areas

Table 5. Potential Losses of Earthquake Disasters

Cluster	Total Potential Loss	Average Potential Loss
0	202558822	1984890
1	98519413	6157463

Table 5 shows that cluster 0 has a greater potential loss value than cluster 1. However, the average of potential loss of, cluster 1 is three times higher to cluster 0. In Table 6 describe average potential losses of earthquake disaster per province and obtain that West Java Province is the region with the highest average potential loss, reaching Rp 160,287,764. Meanwhile, the province with the lowest average potential loss for earthquake disasters is DKI Jakarta, with an average potential loss value of Rp 11,303,088,-. The potential losses referred to in this study refer to economic losses, measured in Rupiah (IDR), as reported by the National Disaster Management Agency (BNPB) from 2012-2022. These losses include damage to infrastructure, collapse of housing, and disruption to economic activities. Damage reports were converted into standardized economic loss estimates to enable comparisons between regions. The estimated potential results for each cluster are an accumulation of potential loss data per cities.

Table 6. Average Potential Losses of Earthquake Disasters Per Province

Province	**Average Potential Loss**
Banten	26761398
DI Yogyakarta	23586596
DKI Jakarta	11303088
West Java	160287764
Central Java	22126570
East Java	56912819

3.2 Clustering of Flood Prone Areas

Through the application of the K Prototype method, it was found that the flood disaster on Java Island could be divided into two clusters ($k = 2$). The determination of this cluster is also supported by the results of the smallest Davis Bouldin index, namely the number of clusters is 2. The following is the division of the cluster:

Table 7. Clustering of Flood Prone Areas

Cluster	**Number of Incidents**	**Number of Casualties**	**Number of Damage to Facilities**
0	20	15366	297
1	65	231600	1095

Table 8. Davis Bouldin Index of Flood

Number of Cluster	Davis Bouldin Index
2	0.886
3	3.585
4	2.350
5	4.064

Table 7 shows that cluster 0 has 79 data and has an average number of events of 20 events, an average number of victims of 15,366 people, and an average number of damages to facilities of 297 units. While in the second cluster (cluster 1) there are 40 data with an average number of flood disasters of 65 events, with an average number of victims of 231,600 people, and an average number of damages to facilities of 1095 units. This means that the areas included in cluster 1 are areas that have a higher level of natural disaster vulnerability than the areas in cluster 0.

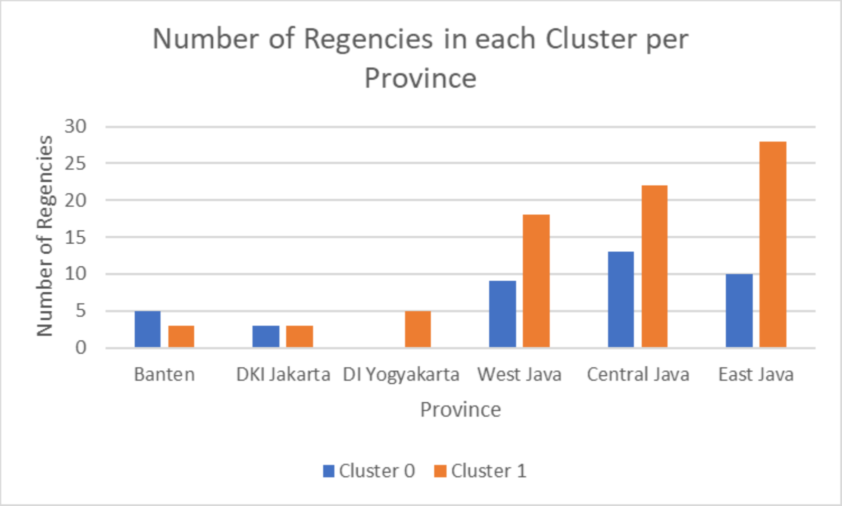

Figure 3. Number of Regencies in each Cluster per Province

Cluster 0 describes flood disasters which tend to have lower values in terms of number of incidents, number of fatalities, and level of damage to facilities. Most cities in the sample are included in this cluster, covering 79 cities (around 66.4% of the total sample). The most common cities or regencies are found in the province of East Java, with 28 cities or regencies followed by Central Java which has 22 cities or regencies in this cluster. However, Banten Province is the province with the highest average number of victims in this cluster with an average of 52,566 people.

Cluster 1 describes flood disasters which tend to have higher values in terms of the number of incidents, number of fatalities, and level of damage to facilities. The number of sample cities in this cluster is 40 cities (around 33.6%). The sample cities in Central Java are the provinces most affected by the flood disaster with 13 cities in this cluster. The provinces of East Java and West Java also have 10 cities and 9 cities that are included in this high-risk cluster. Cluster 1 has an average of 65 flood events per city or district, almost three times more than cluster 0 with low risk. The average number of fatalities in cluster 1 (231600) is much higher than in cluster 0 (15366), while the average damage to facilities in cluster 1 (1095) is almost four times higher than in cluster 0 (297). This shows that cluster 1 has a greater risk and impact from flood disasters compared to cluster 0.

Table 9. The Occurrence of Flood Disaster for Each Province based on Clustering Results

Province	K-Prototypes	Average Occurrence	Number of Incident	Average Number of Casualties	Number of Facility Damages	Average Number of Facility Damages	Number of Cities
Central Java	Cluster 0	24	533	12596	4058	184	22
West Java	Cluster 0	22	393	12636	7683	427	18
East Java	Cluster 0	18	514	16232	11387	407	28
Banten	Cluster 0	12	35	52556	15	5	3
DKI Jakarta	Cluster 0	11	33	30116	3	1	3
DI Yogyakarta	Cluster 0	9	46	1372	354	71	5
Central Java	Cluster 1	74	968	147989	20899	1608	13
West Java	Cluster 1	73	659	459548	10263	1140	9
DKI Jakarta	Cluster 1	62	186	337981	956	319	3
East Java	Cluster 1	60	596	136324	5531	553	10
Banten	Cluster 1	35	173	165405	6150	1230	5

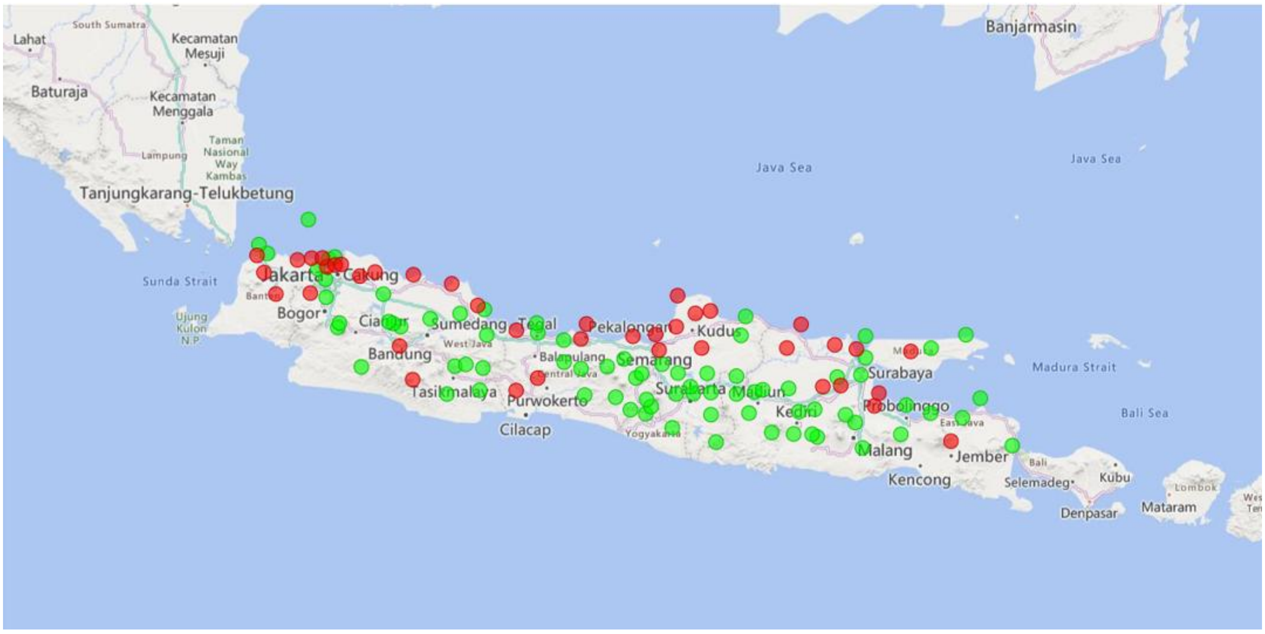

Figure 4. Figure Clustering of Flood Prone Areas

In Fig. 4, red dots represent areas with high risk (cluster 1), while green dots indicate low risk (cluster 0). The areas most affected by disasters are mainly located along the coastal areas facing the Java Sea. This could possibly happen because coastal cities are vulnerable to floods or large tides which cause water overflows to become floods.

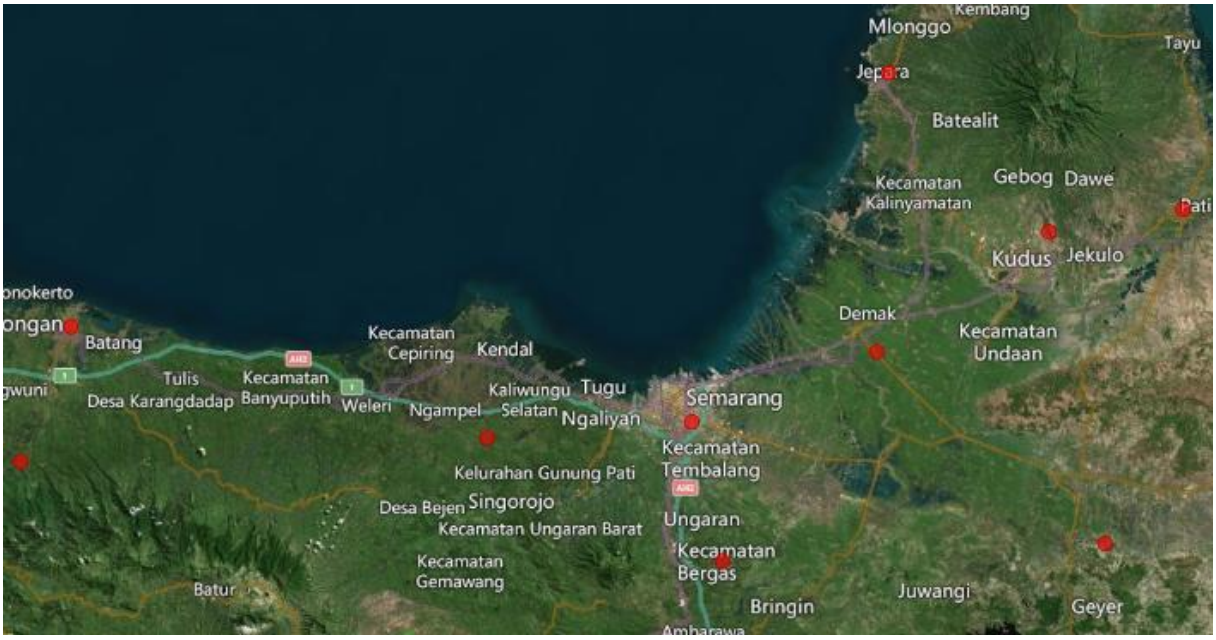

Figure 5. The Zoomed-In Clustering of Flood Prone Areas

Table 10. Potential Losses of Flood Disasters

Cluster	Total Potential Loss	Average Potential Loss
0	107136321	1356156
1	145969674	3649242

The analysis of potential losses resulting from flood disasters indicates that, overall, Cluster 1 exhibits a higher potential loss value than Cluster 0, both in terms of total and average losses, as presented in Table 10. Table 11 describes the average potential flood losses by province and shows that East Java records the highest average potential loss, amounting to IDR 75,078,860. However, this difference is not significantly greater compared to the average losses observed in Central Java and West Java provinces.

Table 11. Average Potential Losses of Flood Disasters Per Province

Province	Average Potential Loss
Banten	23897157
DI Yogyakarta	4362885
DKI Jakarta	29794195
West Java	59318855
Central Java	60654043
East Java	75078860

3.3 Clustering of Tornado Areas

Through the application of the K-Prototypes method, it was found that tornado-prone areas on Java Island could be categorized into two clusters (k = 2). The results of determining the cluster based on the results of the smallest Davis Bouldin index, namely the number of clusters is 2.

Table 12. Davis Bouldin Index of Tornado

Number of Cluster	Davis Bouldin Index
2	1,289
3	1,408
4	1,513
5	1,719

This clustering result provides an overview of the regional distribution based on the frequency of tornado occurrences, the number of casualties, and the extent of material losses, which can serve as a reference for disaster mitigation planning and risk management efforts. In this study, the term "tornado" refers to small-scale local windstorms that occur more frequently in Indonesia, which differ from the large-scale tornadoes commonly found in countries such as the United States. In Indonesia, these local wind disasters are known as Puting Beliung. The following is the division of the cluster:

Table 13. Clustering of Tornado Prone Areas

Cluster	Number of Incidents	Number of Casualties	Number of Damage to Facilities
0	23	229	578
1	109	3545	2523

Cluster 0 describes the Tornado disaster which tends to have lower values in terms of number of incidents, number of fatalities, and level of damage to facilities. Most cities in the sample are included in this cluster, covering 81 of the total 118 cities (around 68.6%). The most common cities or regencies are in the province of East Java, with 31 of the 38 sample cities or regencies located within it. This illustrates that most areas in East Java province tend to experience the Tornado disaster with a relatively lower impact. Cluster 1 describes the Tornado disaster which tends to have higher values in terms of number of incidents, number of fatalities, and level of damage to facilities. The number of sample cities in this cluster is 37 cities (around 31.4%). The sample cities in Central Java are the provinces most affected by the Tornado disaster with 16 cities in this cluster. Cluster 1 had an average of 109 Tornado events per city, almost five times more than cluster 0 with low risk. The average number of fatalities in cluster 1 (3545) is much higher than in cluster 0 (229), while the average damage to facilities in cluster 1 (2523) is almost five times higher than in cluster 0 (578). This shows that cluster 1 has a greater risk and impact from the Tornado disaster compared to cluster 0.

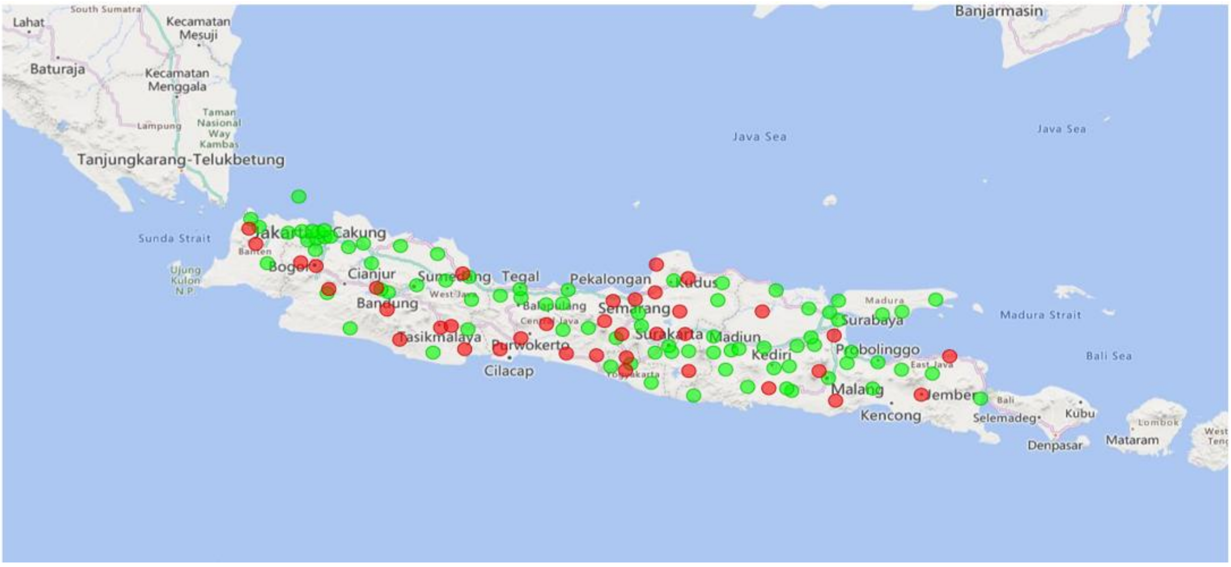

Figure 6. Figure Clustering of Tornado Areas

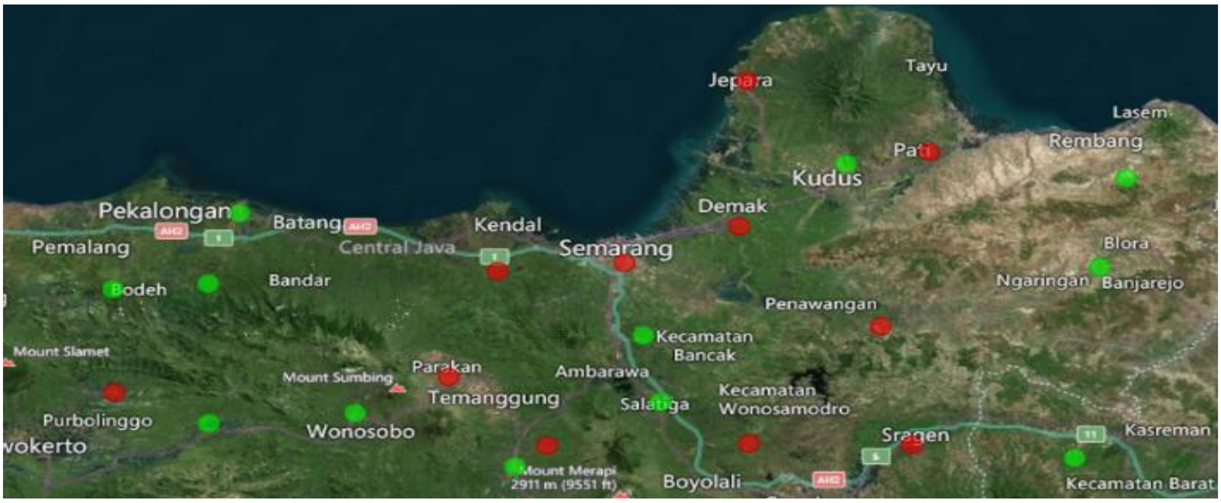

Figure 7. The Zoomed-In Clustering of Tornado Areas

From the cluster map graph, red dots represent areas with high risk (cluster 1), while green dots indicate low risk (cluster 0). The areas most affected by disasters are mainly located along coastal areas facing the Indian Ocean. However, the impact was also seen in the northern part of Central Java Province. This pattern may explain why Central Java has the highest number of tornado disasters on Java Island.

Table 14. Potential Losses of Tornado Disasters

Cluster	Total Potential Loss	Average Potential Loss
0	736406981	9091444
1	417257370	11277226

Analysis of potential losses from the Tornado disaster shows that overall, cluster 0 has a greater potential loss value than cluster 1. However, when averaged, cluster 1 has a higher potential loss than cluster 0. West Java Province is a region with the highest average potential loss, reaching IDR 340,224,460 as shown in Table 15. However, the difference is not very significant with the provinces of East Java and Central Java.

Table 15. Average Potential Losses of Tornado Disasters Per Province

Province	Average Potential Loss
Banten	95169617
DI Yogyakarta	26993157
DKI Jakarta	98417763
West Java	340224460
Central Java	277380567
East Java	315478787

3.4 Clustering of Landslide Prone Areas

Through the application of the K Prototype method, it was found that the landslide disaster on Java Island could be divided into two clusters (k = 2). The results of determining the cluster based on the results of the smallest Davis Bouldin index, namely the number of clusters is 2.

Table 16. Davis Bouldin Index of Landslide

Number of Cluster	Davis Bouldin Index
2	0,777
3	0,910
4	3,514
5	3,203

Table 17. Clustering of Landslide Prone Areas

Cluster	Number of Incidents	Number of Casualties	Number of Damage to Facilities
0	16	163	98
1	134	5066	565

Cluster 0 describes landslide disasters which tend to have lower values in terms of number of incidents, number of fatalities, and level of damage to facilities. The majority of cities in the sample are included in this cluster, covering 83 of the total 118 cities (around 70.3%). The most common cities or regencies are in the province of East Java, with 35 of the 38 sample cities or regencies located within it. However, it can also be seen from Table 17 that the number of damaged facilities in this province is quite high if not averaged out. Cluster 1 describes landslide disasters which tend to have higher values in terms of number of incidents, number of fatalities, and level of damage to facilities. The number of sample cities in this cluster is 35 cities (around 29.7%). The sample of cities or regencies in Central Java is the province most affected by the landslide disaster with 15 cities or regencies followed by West Java Province with 13 cities or regencies. Cluster 1 had an average of 134 landslide events per city or district, more than eight times more than cluster 0 with low risk. The average number of fatalities in cluster 1 (5066) is much higher than in cluster 0 (163), while the average damage to facilities in cluster 1 (565) is almost five times higher than in cluster 0 (98). This shows that cluster 1 has a greater risk and impact from the landslide disaster compared to cluster 0.

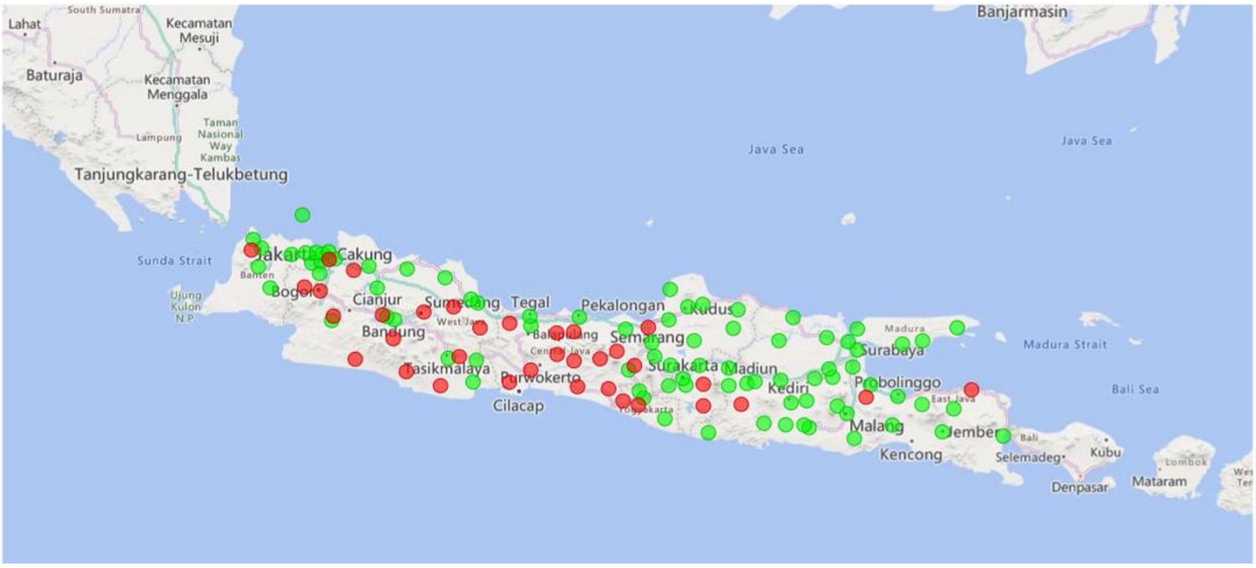

Figure 8. Figure Clustering of Landslide Prone Areas

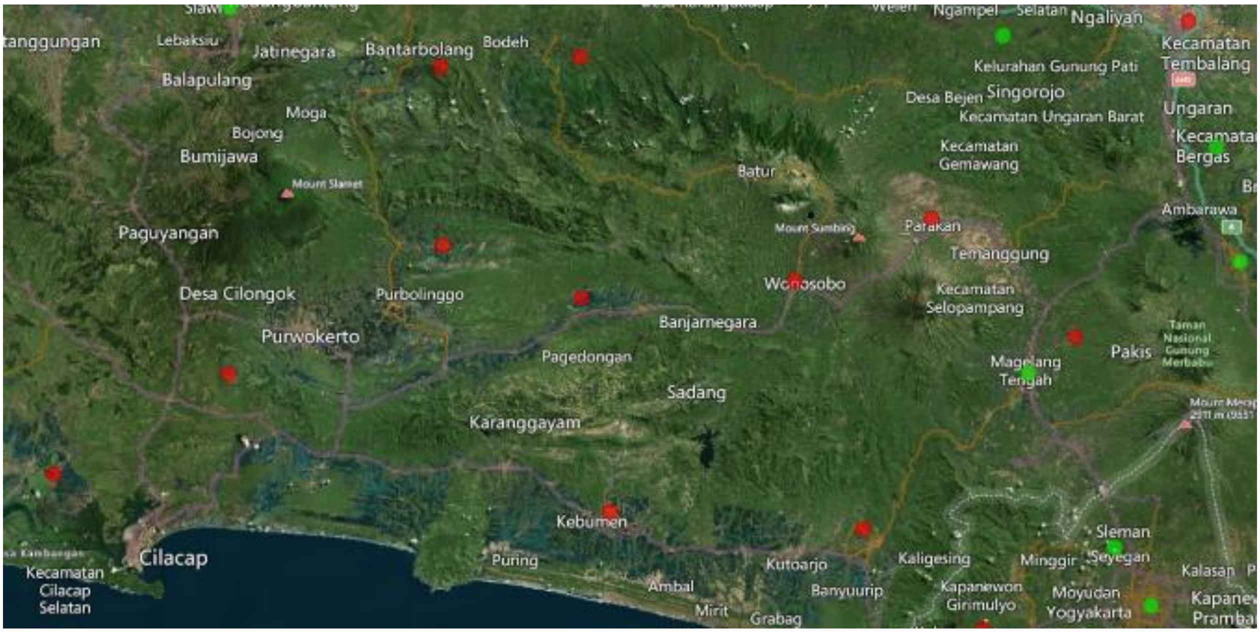

Figure 9. The Zoomed-In Clustering of Landslide Prone Areas

From the cluster map graph, red dots represent areas with high risk (cluster 1), while green dots indicate low risk (cluster 0). The areas most affected by disasters are mainly located in highland areas which increase the risk of landslides.

Table 18. Potential Losses of Landslides Disasters

Cluster	Total Potential Loss	Average Potential Loss
0	21758158	262146
1	63226496	1806471

Based on the Table 18 shows that cluster 1 has a greater potential loss and average potential loss than cluster 0. The potential loss of cluster 1 is almost three times higher to cluster 0 and the average potential loss of cluster 1 has a much higher value than cluster 0.

Table 19. Average Potential Losses of Landslides Per Province

Province	Average Potential Loss
Banten	3127240
DI Yogyakarta	1802631
DKI Jakarta	0
West Java	42792448
Central Java	22496438
East Java	14765896

Analysis of potential losses from the Landslide disaster shows that overall, cluster 1 has a greater potential loss value than cluster 0 both in number and average. In Table 19 show West Java Province is the region with the highest average potential loss, reaching IDR 42,792,448. One of the interesting points is that DKI Jakarta did not experience any potential losses at all due to the landslide disaster.

Based on the overall analysis of disasters, it can be concluded that tornadoes represent the disaster type with the highest level of potential loss and the second most frequent occurrence in Indonesia. On average, each tornado event results in a loss of approximately IDR 200,000.00. Earthquakes rank second in terms of total potential loss among natural disasters. Although earthquakes occur the least frequently among the four disaster types analysed, they cause the highest average potential loss, amounting to approximately IDR 425,000.00 per event. Floods represent the third most frequent disaster and rank third in both total and average potential losses. Meanwhile, landslides are the most frequently occurring natural disaster in Indonesia. However, they contribute the least to total potential losses, resulting in the lowest average loss per incident.

4. CONCLUSION

This research produces clusters of cities potentially prone to earthquakes, floods, Tornadoes, and landslides based on data analysis of natural disaster events in Indonesia from 2012 to 2022. West Java is identified as a region highly vulnerable to all four observed disasters, while DKI Jakarta is an area that is more potentially prone to floods. Central Java shows high vulnerability to floods, landslides, and Tornado, whereas East Java is prone to floods, landslides, and earthquakes. Additionally, based on the analysis of potential losses and clustering results, West Java and East Java are among the top three regions with the highest potential losses for each observed disaster, making them key areas for prioritized mitigation efforts. This research produces clusters of cities that are potentially vulnerable to earthquakes, floods, tornadoes, and landslides based on data analysis of natural disaster events in Indonesia from 2012 to 2022. West Java was identified as a highly vulnerable region to all four observed disasters, while DKI Jakarta Research obtained clustering effectiveness results through the calculation of the Davies-Bouldin Index (DBI). A lower DBI indicates better separation between clusters and higher clustering quality. The DBI obtained indicates that the clusters are reasonably well separated and supports the validity of the clustering results. The findings are expected to be used as future projections to determine the probability of disaster occurrence and estimated losses from disaster probability information, which can then be used for insurance calculations as part of a risk minimisation strategy. Furthermore, the method can be further developed by incorporating additional relevant datasets to enhance analysis accuracy.

Author Contributions

Mey Lista Tauryawati: Conceptualization, Data Curation, Methodology, Software, Visualization, Writing – Original Draft. Ahmad Fuad Zainuddin: Formal analysis, Supervision, Validation, Writing – Review and Editing. Both authors discussed the results and contributed to the final manuscript.

Funding Statement

This research was supported by an internal research grant from Universitas Prasetiya Mulya in 2023.

Acknowledgment

The authors wish to express their appreciation to Prasetiya Mulya University for providing the internal grant and facilities that enabled the successful completion of this research.

Declarations

The authors declare that there is no conflict of interest regarding the publication of this paper. No financial or personal relationships have influenced the work reported in this manuscript.

REFERENCES

[1] Indonesia's National Disaster Management Agency, *GEOPORTAL DATA BENCANA INDONESIA*. Jakarta: Indonesia's National Disaster Management Agency, 2022.

[2] A. F. Pohan *et al.*, "UTILIZATION AND MODELING OF SATELLITE GRAVITY DATA FOR GEOHAZARD ASSESSMENT IN THE YOGYAKARTA AREA OF JAVA ISLAND, INDONESIA," *Kuwait J. Sci.*, vol. 50, no. 4, pp. 499–511, Oct. 2023, doi: https://doi.org/10.1016/j.kjs.2023.05.016.

[3] PUSGEN, "PETA BAHAYA GEMPA INDONESIA," Pusat Studi Gempa Nasional, 2017.

[4] R. Muliono *et al.*, "DATA MINING CLUSTERING MENGGUNAKAN ALGORITMA K-MEANS UNTUK KLASTERISASI TINGKAT TRIDHARMA PENGAJARAN DOSEN," *CESS J. Comput. Eng. Syst. Sci.*, vol. 4, no. 2, pp. 272–279, 2019.

[5] M. Bennett, E. J. Kleczyk, K. Hayes, and R. Mehta, "EVALUATING SIMILARITIES AND DIFFERENCES BETWEEN MACHINE LEARNING AND TRADITIONAL STATISTICAL MODELING IN HEALTHCARE ANALYTICS," in *Artificial Intelligence*, vol. 12, M. Antonio Aceves Fernandez and C. M. Travieso-Gonzalez, Eds., IntechOpen, 2022. doi: https://doi.org/10.5772/intechopen.105116.

[6] S. Annas, I. Irwan, R. H. Safei, and Z. Rais, "K-PROTOTYPES ALGORITHM FOR CLUSTERING THE TECTONIC EARTHQUAKE IN SULAWESI ISLAND," *J. Varian*, vol. 5, no. 2, pp. 191–198, May 2022, doi: https://doi.org/10.30812/varian.v5i2.1908.

[7] M. Refaldy, S. Annas, and Z. Rais, "K-PROTOTYPE ALGORITHM IN GROUPING REGENCY/CITY IN SOUTH SULAWESI PROVINCE BASED ON 2020 PEOPLE'S WELFARE," *ARRUS J. Math. Appl. Sci.*, vol. 3, no. 1, pp. 11–19, May 2023, doi: https://doi.org/10.35877/mathscience1763.

[8] O. U. Mehmood, Y. B. Wah, and W. F. Wan Yaacob, Eds., *DECISION MATHEMATICS, STATISTICAL LEARNING AND DATA MINING: SELECTED CONTRIBUTIONS FROM ICMSCT2023, Manila, Philippines, September 20-21*. in Springer Proceedings in Mathematics & Statistics, vol. 461. Singapore: Springer Nature Singapore, 2024.

[9] M. A. Bramer, *PRINCIPLES OF DATA MINING*, 3rd ed. 2016. in Undergraduate Topics in Computer Science. London: Springer, 2016. doi: https://doi.org/10.1007/978-1-4471-7307-6.

[10] P.-N. Tan, M. Steinbach, A. Karpatne, and V. Kumar, *INTRODUCTION TO DATA MINING*, Second edition, Global edition. Harlow: Pearson Education, 2019.

[11] E. Muningsih and S. Kiswati, "PENERAPAN METODE K-MEANS UNTUK CLUSTERING PRODUK ONLINE SHOP DALAM PENENTUAN STOK BARANG," *Bianglala Inform.*, vol. 3, no. 1, pp. 1–11, 2015.

[12] P. Fränti and S. Sieranoja, "HOW MUCH CAN K-MEANS BE IMPROVED BY USING BETTER INITIALIZATION AND REPEATS?," *Pattern Recognit.*, vol. 93, pp. 95–112, Sep. 2019, doi: https://doi.org/10.1016/j.patcog.2019.04.014.

[13] K. M. P. Dewi, M. L. Tauryawati, and A. F. Zainuddin, "CLUSTERING OF DISASTER PRONES AREAS IN JAVA ISLAND," presented at the RECENT ADVANCES IN MATERIALS AND MANUFACTURING: ICRAMM2023, Erode, India, 2024, p. 020003. doi: https://doi.org/10.1063/5.0230700.

[14] STMIK STIKOM Indonesia, D. A. I. C. Dewi, D. A. K. Pramita, and STMIK STIKOM Indonesia, "ANALISIS PERBANDINGAN METODE ELBOW DAN SILHOUETTE PADA ALGORITMA CLUSTERING K-MEDOIDS DALAM PENGELOMPOKAN PRODUKSI KERAJINAN BALI," *Matrix J. Manaj. Teknol. Dan Inform.*, vol. 9, no. 3, pp. 102–109, Nov. 2019, doi: https://doi.org/10.31940/matrix.v9i3.1662.

[15] S. Aisah, I. Aknuranda, and A. N. Rusydi, "SISTEM PENDUKUNG KEPUTUSAN UNTUK PENGELOMPOKAN BARANG TERJUAL PADA PT DASEMA DIGI PERSADA DENGAN METODE K-MEANS CLUSTERING," *J. Pengemb. Teknol. Inf. Dan Ilmu Komput.*, vol. 4, no. 7, pp. 2309–2317, 2020.

[16] G. Vardakas, I. Papakostas, and A. Likas, "DEEP CLUSTERING USING THE SOFT SILHOUETTE SCORE: TOWARDS COMPACT AND WELL-SEPARATED CLUSTERS," 2024, *arXiv*. doi: http://doi.org/10.48550/ARXIV.2402.00608.

[17] M. Gagolewski, M. Bartoszuk, and A. Cena, “ARE CLUSTER VALIDITY MEASURES (IN) VALID?,” *Inf. Sci.*, vol. 581, pp. 620–636, Dec. 2021, doi: https://doi.org/10.1016/j.ins.2021.10.004.

[18] D. L. Davies and D. W. Bouldin, “A CLUSTER SEPARATION MEASURE,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-1, no. 2, pp. 224–227, Apr. 1979, doi: https://doi.org/10.1109/TPAMI.1979.4766909.

[19] A. Hamadani *et al.*, “OUTLIER REMOVAL IN SHEEP FARM DATASETS USING WINSORIZATION,” *Bhartiya Krishi Anusandhan Patrika*, no. Of, Jan. 2022, doi: https://doi.org/10.18805/BKAP397.

[20] Abuzaid, “A COMPARATIVE STUDY ON UNIVARIATE OUTLIER WINSORIZATION METHODS IN DATA SCIENCE CONTEXT,” *Stat. Appl. – Ital. J. Appl. Stat.*, vol. 36, no. 1, p. 1, 2024, doi: http://doi.org/10.26398/IJAS.0036-004.